Statistical Machine Learning (BE4M33SSU)
Lecture 3: Empirical Risk Minimization

Czech Technical University in Prague
V. Franc

Learning

- **The goal:** Find a strategy $h\colon \mathcal{X} \to \mathcal{Y}$ minimizing $R(h)$ using the training set of examples

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$$

drawn from i.i.d. according to unknown $p(x, y)$.

- **Hypothesis class:**

$$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$$

- **Learning algorithm:** a function

$$A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$$

which returns a strategy $h_m = A(\mathcal{T}^m)$ for a training set $\mathcal{T}^m$

Learning: Empirical Risk Minimization approach

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from the training examples $\mathcal{T}^m$,

$$R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y^i, h(x^i))$$

- The ERM based algorithm returns h_m such that

$$h_m \in \operatorname*{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h) \tag{1}$$

- Depending on the choince of $\mathcal{H}$ and ℓ and algorithm solving (1) we get individual instances e.g. Support Vector Machines, Linear Regression, Logistic Regression, Neural Networks learned by back-propagation, AdaBoost, Gradient Boosted Trees, ...

Excess error = Estimation error + Approximation errors

The characters of the play:

- $R^* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R(h)$ best attainable true risk
- $R(h_\mathcal{H})$ best risk in $\mathcal{H}$ where $h_\mathcal{H} \in \operatorname{Argmin}_{h \in \mathcal{H}} R(h)$
- $R(h_m)$ risk of $h_m = A(\mathcal{T}_m)$ learned from $\mathcal{T}^m$

Excess error: the quantity we want to minimize

$$\underbrace{\Big(R(h_m) - R^*\Big)}_{\text{excess error}} = \underbrace{\Big(R(h_m) - R(h_\mathcal{H})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(h_\mathcal{H}) - R^*\Big)}_{\text{approximation error}}$$

Questions:

- Which of the quantities are random and which are not ?
- What causes individual errors ?
- How do the errors depend on $\mathcal{H}$ and m?

Statistically consistent learning algorithm

- The statistically consistent algorithm can make the estimation error $R(h_m) - R(h_{\mathcal{H}})$ arbitrarily small if it has enough examples.
- Is the ERM algorithm statistically consistent ?

Definition 1. *The algorithm $A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ is statistically consistent in $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if for any $p(x,y)$ and $\varepsilon > 0$ it holds that*

$$\lim_{m\to\infty} \mathbb{P}\bigg(R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \bigg) = 0$$

where $h_m = A(\mathcal{T}^m)$ is the hypothesis returned by the algorithm A for training set $\mathcal{T}^m$ generated from $p(x,y)$.

Example: ERM does not work if $\mathcal{H}$ is unconstrained

- Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$.

- The optimal strategy is $h(x) = +1$ with the Bayes risk $R^* = 0.2$.

- Consider learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \dots, (x^m, y^m)\}$ returns strategy

$$h_m(x) = \begin{cases} y^j & \text{if } \; x = x^j \; \text{ for some } \; j \in \{1, \dots, m\} \\ -1 & \text{otherwise} \end{cases}$$

- The empirical risk is $R_{\mathcal{T}^m}(h_m) = 0$ with probability 1 for any m.

- The expected risk is $R(h_m) = 0.8$ for any m.

Uniform Law of Large Numbers

- We say that training set $\mathcal{T}^m$ is "bad" for $h \in \mathcal{H}$ if the generalization error is $|R(h) - R_{\mathcal{T}^m}(h)| \geq \varepsilon$.
- ULLN holds for $\mathcal{H}$ provided the probability of seeing at least one "bad training set" can be made arbitrarily low if we have enough examples.

Definition 2. *The hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ satisfies the uniform law of large numbers if for all $\varepsilon > 0$ and $p(x,y)$ generating $\mathcal{T}^m$ it holds that*

$$\lim_{m \to \infty} \mathbb{P}\left(\sup_{h \in \mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big| \geq \varepsilon \right) = 0$$

Theorem 1. *If $\mathcal{H}$ satisfies ULLN then ERM is statistically consistent in $\mathcal{H}$.*

ULLN for finite hypothesis class

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- Define the set of all “bad” training sets for a strategy $h \in \mathcal{H}$ as

$$\mathcal{B}(h) = \Big\{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \,\Big|\, \big| R_{\mathcal{T}^m}(h) - R(h) \big| \geq \varepsilon \Big\}$$

- Hoeffding inequality generalized for finite hypothesis class $\mathcal{H}$:

$$\mathbb{P}\Big(\max_{h \in \mathcal{H}} \big| R_{\mathcal{T}^m}(h) - R(h) \big| \geq \varepsilon \Big) \leq \sum_{h \in \mathcal{H}} \mathbb{P}\big(\mathcal{T}^m \in \mathcal{B}(h) \big)\big) = 2\,|\mathcal{H}|\, e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$$

- Therefore

$$\lim_{m \to \infty} \mathbb{P}\Big(\max_{h \in \mathcal{H}} \big| R_{\mathcal{T}^m}(h) - R(h) \big| \geq \varepsilon \Big) = 0$$

Corrollary 1. *The ULLN is satisfied for a finite hypothesis class.*

Proof: ULLN implies consistency of ERM

For fixed $\mathcal{T}^m$ and $h_m \in \operatorname{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h)$ we have:

$$
\begin{aligned}
R(h_m) - R(h_{\mathcal{H}}) &= \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_m) - R(h_{\mathcal{H}})\Big) \\
&\leq \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_{\mathcal{H}}) - R(h_{\mathcal{H}})\Big) \\
&\leq 2 \sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big|
\end{aligned}
$$

Therefore $\varepsilon \leq R(h_m) - R(h_{\mathcal{H}})$ implies $\frac{\varepsilon}{2} \leq \sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big|$ and

$$
\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \Big) \leq \mathbb{P}\Big(\sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big| \geq \frac{\varepsilon}{2} \Big)
$$

so if converges the RHS to zero (ULLN) so does the LHS (estimation error).

Linear classifier minimizing classification error

- $\mathcal{X}$ is a set of observations and $\mathcal{Y} = \{+1, -1\}$ a set of hidden labels
- $\phi\colon \mathcal{X} \to \mathbb{R}^n$ is fixed feature map embedding $\mathcal{X}$ to $\mathbb{R}^n$
- **Task:** find linear classification strategy $h\colon \mathcal{X} \to \mathcal{Y}$

$$h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b) = \begin{cases} +1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b \geq 0 \\ -1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b < 0 \end{cases}$$

with minimal expected risk

$$R^{0/1}(h) = \mathbb{E}_{(x,y)\sim p}\Big(\ell^{0/1}(y, h(x))\Big) \quad \text{where} \quad \ell^{0/1}(y, y') = [\![y \neq y']\!]$$

- We are given a set of training examples

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$$

drawn from i.i.d. with the distribution $p(x, y)$.

ERM learning for linear classifiers

- The Empirical Risk Minimization principle leads to solving

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n\times\mathbb{R})} R^{0/1}_{\mathcal{T}^m}(h(\cdot;\boldsymbol{w},b)) \qquad (1)$$

  where the empirical risk is

$$R^{0/1}_{\mathcal{T}^m}(h(\cdot;\boldsymbol{w},b)) = \frac{1}{m}\sum_{i=1}^{m}[\![y^i \neq h(x^i;\boldsymbol{w},b)]\!]$$

We will address the following issues:

1. The statistical consitency of the ERM for hypothesis class $\mathcal{H} = \{h(x) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b) \mid (\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}\}$.

2. Algorithmic issues (next lecture): in general, there is no known algorithm solving the task (1) in time polynomial in m.

Vapnik-Chervonenkis (VC) dimension

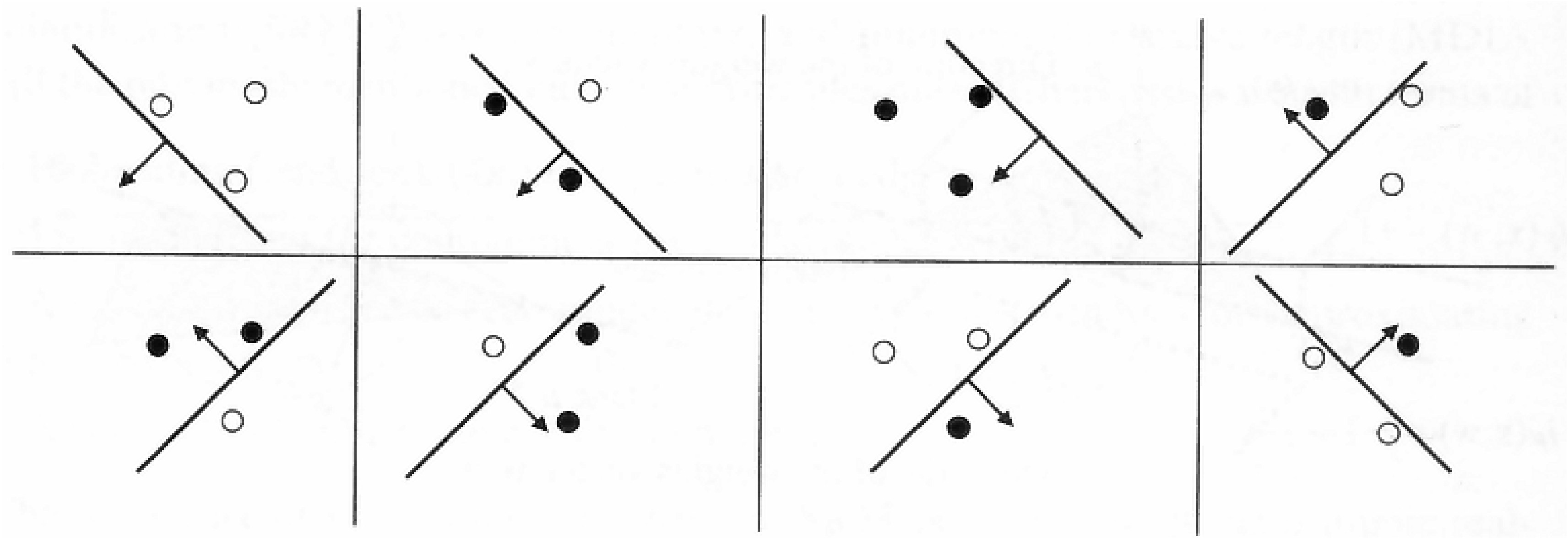
Definition 3. *Let* $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ *and* $\{x^1,\ldots,x^m\} \in \mathcal{X}^m$ *be a set of* m *input observations. The set* $\{x^1,\ldots,x^m\}$ *is said to be shattered by* $\mathcal{H}$ *if for all* $\boldsymbol{y} \in \{+1,-1\}^m$ *there exists* $h \in \mathcal{H}$ *such that* $h(x^i) = y^i$, $i \in \{1,\ldots,m\}$.

Definition 4. *Let* $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$. *The Vapnik-Chervonenkis dimension of* $\mathcal{H}$ *is the cardinality of the largest set of points from* $\mathcal{X}$ *which can be shattered by* $\mathcal{H}$.

VC dimension of class of two-class linear classifiers

Theorem 2. *The VC-dimension of the hypothesis class of all two-class linear classifiers operating in n-dimensional feature space* $\mathcal{H} = \{h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b) \mid (\boldsymbol{w}, b) \in (\mathbb{R}^n \times \mathbb{R})\}$ *is* $n+1$.

Example for $n = 2$-dimensional feature class



Consistency of prediction with two classes and $0/1$-loss

Theorem 3. *Let $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d < \infty$ and $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ a training set draw from i.i.d. rand vars with distribution $p(x, y)$. Then, for any $\varepsilon > 0$ it holds*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 4 \left(\frac{2\, e\, m}{d} \right)^d e^{-\frac{m\, \varepsilon^2}{8}}$$

Corollary 1. *Let $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d < \infty$. Then ULLN applies and hence ERM is statistically consistent in $\mathcal{H}$ w.r.t $\ell^{0/1}$ loss function.*